

DOCUMENT RESUME

ED 290 159

CS 211 026

AUTHOR Deluzain, Edward
TITLE The Use of Holistic Scoring in Program Evaluation.
PUB DATE Mar 84
NOTE 13p.; Paper presented at the Annual Meeting of the Conference on Writing Assessment (2nd, Tallahassee, FL, March 7-9, 1984). For a related document, see CS 211 027.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Educational Administration; Educational Diagnosis; *Evaluation Methods; High Schools; *Holistic Evaluation; *Program Evaluation; Scoring; Skill Analysis; Student Evaluation; Writing (Composition); Writing Evaluation; Writing Research
IDENTIFIERS Florida

ABSTRACT

In addition to determining students' writing proficiency, holistic writing scores obtained through a standardized evaluation provide useful data in making value judgments concerning various aspects of high school writing programs. In a cross-sectional study using holistic scoring, average scores of groups of students can be compared, and the results can identify strengths and weaknesses in course content and standards set for students by teachers. At Mosley High School in Florida, for example, results indicated that the writing performance of students enrolled in writing courses was lower than the performance of students in literature courses. After analyzing these results, a more extensive reading program was instituted in the writing courses in order to improve students' reading, thinking, analyzing and interpretive skills. However, when the scores for accelerated students indicated no greater rate of improvement between grades than for the other tracks (regular and basic), teachers changed their expectations for accelerated students, making the courses more demanding. Complete objectivity is necessary for this type of program evaluation, and caution is needed when choosing outside evaluators, to insure accurate assessment. (MM)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

The Use of Holistic Scoring in Program Evaluation

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

Edward Deluzain

Edward Deluzain, Ph.D., Chair
Department of English and Foreign Languages

Mosley High School
3418 North Palo Alto Avenue
Panama City, FL 32405

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☐ This document has been reproduced as
received from the person or organization
originating it.
☐ Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official
OERI position or policy.

ED290159

Holistic scoring is gaining popularity in a variety of contexts within the educational community, and it is being used for a variety of purposes, including determining whether students have achieved "competency" in writing. In Florida, for example, college students are required to pass the College-level Academic Skills Test (CLAST) before they are allowed to move from lower-division to upper-division status in state universities. One component of the CLAST is an essay written under standardized conditions and scored holistically on a four-point scale. One criterion for passing the exam is achieving a passing score on the essay portion. The high validity of holistic scoring and the relative ease with which the technique can be used will probably make it even more popular as a means of evaluating students in the future.

Another use of holistic scoring which has not been explored to any great extent is in the area of the evaluation of whole programs in English, such as the English curriculum of a high school. For the last two years at Mosley High School in Panama City, Florida, I have been working on a system for evaluating our writing curriculum in grades 10, 11, and 12, using holistic scoring of compositions written under standardized conditions and scored by the method outlined by Miles Myers in A Procedure for Writing Assessment and Holistic Scoring (Urbana, Ill.: ERIC Clearinghouse on Reading and Communication Skills and the National Council of Teachers of English, 1980).

CS 211026

My experience with scoring of the papers and with the analysis of the scores leads me to believe that holistic scoring offers us a valuable tool to use in program evaluation in the area of writing.

The behavioral objectives movement of the late 1960's and early 1970's was an attempt at program evaluation, but it failed to achieve any meaningful results in the area of writing because it began by asking us to describe the end. The premise underlying behavioral objectives is that it is possible to spell out in terms of specific behaviors what it is that teachers expect students to be able to do after a course of instruction. As conceived by Robert F. Mager (Preparing Instructional Objectives. Palo Alto, Calif.: Fearon Publishers, Inc., 1962) and others, it seems reasonable to expect a shop teacher, let's say, to know the specifications of an "effective" doghouse, to be able to describe the doghouse very precisely, and to analyze the tasks necessary to produce it, given hammer, nails, wood, ... This, of course, assumes two things which may be true of effective doghouses but which are decidedly not true of effectively paragraphs: first, that all effective doghouses look pretty much like one another; and, second, that the steps in building one are pretty much the same for everybody. In applying these assumptions to writing, the argument went something like this: If these assumptions are true for building doghouses (incidentally, Mager drew examples to support the reasonableness of behavioral objectives from industrial arts), then common sense tells us that they must be equally true for building paragraphs and essays. In this case, as in so many others, Stuart Chase's definition of common sense as that which tells us that the world is flat once again holds true. English teachers rightly resisted the behavioral objectives approach to their subject, but one unfortunate side effect was that program evaluation, of which behavioral objectives are the servants and not the master, lost

ground because everyone concluded that it couldn't be done without oversimplifying and trivializing the discipline.

In its simplest and most elegant form, program evaluation consists of a three-step process: 1.) spelling out what you want to accomplish; 2.) using some method or program of instruction to accomplish what you say you want to accomplish; and, 3.) comparing the final accomplishment to what you set out to accomplish in the first place. The model in this form doesn't work in evaluating an instructional program in writing because the model, like behavioral objectives, is based on the assumptions that it is possible to specify ahead of time in precise terms the desired outcome of instruction and that the desired outcome is possible to achieve. Like the person who takes up weight lifting to become stronger but who does not have a precise idea of what "stronger" means, we want our students' writing to improve, but we don't know what we can reasonably expect to find in pieces of writing produced after x-amount of instruction. For that matter, we also don't know the limits of what it is possible for students to achieve. Is Shakespeare the ultimate? Eliot? Faulkner? Perhaps research will some day give us an answer to the question of reasonable expectations, but we will certainly never know the limits of the possible.

Since we cannot now describe goals for writing instruction in the way required by the structure of behavioral objectives, and since program evaluation depends on a comparison of expectations and results, can we ever hope to really evaluate what we're doing in the teaching of writing? I, of course, think we can, and I think holistic scoring offers a methodology that we can use to do this. In order to use this methodology for program evaluation, we have to content ourselves with goals which are reasonable and realistic in light of the limitations in the state of the knowledge

about the teaching and evaluating of writing. One such goal is that graduating seniors should write better than entering sophomores.

One way to determine whether this goal is met in a particular school program is to have entering sophomores write a composition on a specified topic under standardized conditions and to have them write the same composition under the same conditions shortly before they graduate three years later. The pairs of papers could then be judged anonymously without dates by readers who are told to pick the better of the two. If the paper written in the senior year is ^k _^ piced, then the conclusion that the student's writing has improved is justified. If this happens in a large enough number of cases, it would be possible to infer that, on the average, the writing program of the school is effective in improving students' writing.* This type of longitudinal evaluation system would work, but it would be so cumbersome and the results would be so long in coming that it probably would not be practical in most school settings.

Another way to obtain the same information is through a cross-sectional study using holistic scoring to produce data we can trust. This is the type of study we use at Mosley High School. We ask students in all three grades to write compositions on the same topic, on the same day, during their English classes. The papers are coded with numbers to preserve anonymity and are then thoroughly randomized. Each paper is scored holistically on

*In any type of evaluation study of improvement in writing we have to be very careful of drawing conclusions that go beyond the state of the knowledge about how much development would occur naturally without instruction or practice. As far as I know, no research has ever been undertaken to determine this, and research of this type would probably raise serious ethical questions. However, without this type of research we can't say with certainty that improvement over time is due to instruction and not to maturation. While we can assume that instruction affects performance positively, we have to beware of the post hoc, ergo propter hoc fallacy.

a six-point scale by two teachers (and by a third teacher if the scores from the first two are more than one point apart), and the two scores are added together to produce summary scores which range from 2 to 12. We preserve the names of the students and information about their English courses on index cards which are also coded, so we are able to compare average scores of groups of students.

In both of our assessments^s, the mean scores of seniors have been higher than the mean scores of sophomores, and the differences between the means have been statistically significant at the .01 level. This finding may also be true of every school in the United States, but then again, maybe not. My point is that we now have hard scientific data to prove to anyone who asks that our students improve in their writing while they're in our program, and there are probably very few schools which can make that claim. We have answered the basic accountability question within the limitations of the state of the knowledge, and we have done so without a single behavioral objective and with the investment of only two days' time each year (one day to write the compositions and one day to score them). When I remember the hours I spent trying to write behavioral objectives, and the even greater number of hours I spent arguing that behavioral objectives in writing were profanations before the muses, I am truly awed by what holistic scoring has enabled us to accomplish.

An evaluation system of this type, which confirms what is assumed and which proves that growth does occur, is very valuable in boosting morale of both students and teachers, and it can have considerable value in marshalling support for a program. However, in my opinion, the real value of the evaluation system we use at Mosley lies in its ability to identify areas of strength and

weakness within our program and to suggest ways in which the program can be improved. I doubt that a longitudinal plan like the one I outlined earlier could accomplish this effectively. I will describe two ways in which our writing assessment program enabled us to spot areas of weakness as examples of how a cross-sectional evaluation study can accomplish this.

My first example concerns course content. All of our students are required to take a one-semester course in literature and a one-semester course in writing each year to satisfy their English requirement for graduation. The courses are taken in random order.

With this kind of curriculum, it is reasonable to assume that students' writing improves more in the writing courses than it does in the literature courses, although all students are assigned one writing task each week in all courses. By analyzing data from our first writing assessment by type of course, we were able to determine that this was, in fact, a false assumption. We administer the writing assessment at the start of the second semester, and the first year we found that the average scores were higher for students who had just finished a literature course in 66% of our comparison groups than for those who had just finished a writing course. When I first realized this, I thought that somehow the brighter students had ended up by chance in literature courses first semester, but when I compared mean scores on a standardized achievement test for the groups involved, I found no differences that were in any way significant. This then led to a logical examination of possible causes, and I realized that even though there was little or no formal instruction in writing strategies in the literature courses, the reading and discussing the students did in those courses actually constituted more and better pre-writing activity than time allowed to occur in the writing

courses. In other words, I concluded that the reading, talking, thinking, analyzing, and interpreting the students did in literature contributed to their being better writers than did the direct instruction in rhetorical strategies and editing skills they were receiving in their writing courses.

On the basis of this conclusion we instituted a more extensive reading program in our writing courses at the start of this school year. Students are now required to read novels and long plays in their writing courses, and these works are taught in the same way as they are taught in literature courses. In our second writing assessment, which we conducted at the start of the second semester of this school year, the students who had literature first semester had higher mean scores in only 37.5% of the comparison groups. Since the addition of the reading requirements was the only change we made in the structure of the writing courses, I feel justified in concluding that the addition of the literature study is responsible for making the writing courses more effective in actually producing higher writing scores for a greater number of students. This is not a conclusion verified through controlled research, of course, and there may be other factors responsible for the change, but short of undertaking a study to investigate the matter more scientifically, I am satisfied that my original conclusion about the effectiveness of literature study as a means of improving writing was correct.

It might well be argued that the kind of change in our program we made was logical and the need for the change should have been apparent without having to go through a writing assessment to arrive at the same conclusions. This argument is not without merit, but it overlooks several important factors. First, there is the question of how we would have ever come to the conclusion that the writing courses were less effective in bringing

about improvement in writing than the literature courses. At the time of the first assessment, I was teaching two sections of a one-semester writing course to juniors and one section of a year-long AP English course to seniors. I was convinced that I was really teaching my students how to write, and I firmly believed that my students from the first semester would do much better on the writing assessment than the students in the literature courses. The results were one of the major surprises of my professional life, and my limited vision, like the limited vision of everyone else in the department, was inadequate to allow me to see the true picture within the total program. Second, our curriculum was well established, and even if I had been able to perceive the need for a major change, I doubt that the other teachers in the department would have been willing to go along with the change just because I suspected problems might exist. The data, though, were irrefutable, especially since all of the teachers in the department had been involved in administering the assessment and in scoring the papers. Third, the addition of major reading requirements in the writing courses required the expenditure of several thousands of dollars to support it. I am convinced that my principal took my request for additional money more seriously when I approached him with data to support my request than if I had simply asked for the money because I thought the change sounded like a good idea.

My second example concerns standards and expectations set for students by teachers. Our students are grouped by ability into three tracks: basic, average, and accelerated. I had reasoned that students in accelerated classes should be expected to have the highest mean scores for their grade levels, and this was the case at every grade. However, I also reasoned that the rate of improvement over three years should be greater and should increase, or at least remain constant, for accelerated students. My analysis of the data from

the first assessment showed that this was not the case.

In order to determine rate, I calculated means and standard deviations for all students in each grade and then I found means for each track within grades. I subtracted the mean for the track from the mean for the grade, and I divided the difference by the standard deviation for the grade. This provided a measure of the relative standing of each track with respect to the mean for the grade. I expected to find that the accelerated students increased their standing with respect to their grade as they proceeded from sophomore to senior year. Instead, I observed just the opposite. The accelerated sophomores had relatively better scores (compared to all sophomores) than the accelerated juniors (compared to all juniors), and the accelerated juniors had better scores than the accelerated seniors (compared to all seniors). This indicated that the rate of growth with respect to all members of their respective classes declined for accelerated students from the sophomore year to the senior year.

I had suspected that we were probably not demanding enough of our most able students, and these data confirmed that suspicion. The only intervention that occurred was to try to heighten the awareness of the teachers to the need to have tougher standards for the accelerated students, who by and large are less resistive to increased demands and who were already doing better than any other group. The teachers responded by changing their expectations for students in their classes, and these changes had the effect, in my opinion, of making the courses more demanding. The results of this year's writing assessment indicate that these changes were sufficient to make a difference with our current accelerated seniors. I calculated the "rate" statistic on the old scores for students who participated in both assessments

and compared these figures for the identical students this year. The results showed that the accelerated juniors (who were accelerated sophomores last year) experienced a slight decline in their standing with respect to other juniors (from +1.142 to +1.131), but the accelerated seniors this year made a slight advance in their relative standing (from +0.765 to +0.975). Next year we will be able to look at the rate of growth of the senior class over three years, and this comparison should provide insight into their decline in relative standing this year as juniors.

My point throughout this discussion has been that the holistic writing scores we obtained through a standardized writing assessment have been strong, clear data for us to use in making value judgments about various aspects of our writing program. I know of no standardized test which could have provided us with anything close to the kind of information we obtained by actually having students write compositions. We have been able to draw conclusions and infer causes, and the solutions we developed to solve problems are apparently having some effect on our students' writing. We have retained all of the compositions from both assessments (about 2,200 papers in all), and when time permits I plan to study various aspects of writing, such as syntactic maturity as measured by t-units and the number and types of errors our students make, and I will compare our results to national normative data from the National Assessment of Educational Progress and other sources. Then, perhaps, we will really know whether our students' writing is as mature as it possibly could be for their ages and whether their errors are developmentally consistent or a real cause for concern.

In my analyses of the data from our writing assessments I have used statistical tests of significance, and we have tried to standardize our

procedures as much as is possible in the real-life setting of a high school. I have not tried to conduct tightly controlled research based on a control group-experimental group model. The reason for this is quite simple: we were interested in developing a process to find out about our own program and about how our students perform in writing; we were not interested in adding to the knowledge about writing. This is not to denigrate research, of course. Research results are the life-blood of program evaluation because they tell us which goals are realistic and which sets of results indicate problems. However, the kind of research in writing we need right now is beyond the scope of most high school English departments acting on their own. On the other hand, program evaluation studies are most accessible.

A few last words about program evaluation in general are in order.

First, the term "program evaluation" is often used by state legislatures, school boards, and other governing bodies without much precision. I get the impression that these groups think program evaluation is easy to accomplish and that it is the kind of thing every principal and teacher could do without much effort if they would just put down their coffee cups long enough to do it. Consequently, these bodies rarely provide sufficient funds to hire evaluators, to develop instruments for data collection where none exist, to analyze data, to write reports, etc. A good illustration of this is the fact that Florida is spending in excess of \$20 million this year alone to implement the writing program mandated by the Jack Gordon Writing Skills Enhancement Act, but not a single dime is being spent to evaluate it. Old saws about squeezing blood from turnips and about getting what you pay for come to mind.

Second, certain types of program evaluation studies should never be undertaken by people who have a personal stake in the outcome. We undertook an evaluation of our writing program at Mosley because we wanted to, not

because we had to. We certainly weren't going to stop teaching writing if we had discovered that our students didn't make gains from their sophomore year to their senior year, and no writing teacher's job was on the line when we discovered that students taking literature wrote better than those taking writing. Even so, we did not approach the task with complete objectivity. I think we mined the data for all the nuggets, and I believe we drew correct inferences from the results. But I don't know these things, and I also don't know how an outside evaluator would have interpreted our scores. In-house evaluation can be very helpful for program improvement, as I think I've demonstrated, but for final "GO/NO GO" decisions, objective outsiders should be consulted, and this consultation should begin before the first goal is set and before the first lesson is taught. A word of caution about outside evaluators is needed, too. A person with impeccable credentials in evaluation may be no more knowledgeable about a program--such as writing instruction--than teachers are about program evaluation. If an evaluator doesn't know the right questions to ask, how can anyone trust the answers?

Presented at The Second Annual Conference on Writing Assessment
 March 7, 8, and 9, 1984
 Tallahassee, Florida